LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-607577

Null Hypothesis Significance Testing for Trace Chemical Weapon Analyte Detection

S. P. Velsko

December 10, 2012

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-TR-607577

Null Hypothesis Significance Testing for Trace Chemical Weapon Analyte Detection in Forensic Contexts

Steve Velsko
Lawrence Livermore National Laboratory
Livermore, California

December 11, 2012

Abstract

This report outlines an approach to statistical significance testing for trace analyte detection in the context of forensic investigations of chemical weapons related activities. Simple expressions for P values under null hypotheses for background and contamination are derived. An explicit rationale for choosing an α value is outlined, and some practical issues for implementing defensible null hypothesis significance testing in the CW forensics context are discussed. There are a number of limitations and precautions that must be appreciated for effective use of this method. Statistical significance testing has an important role in integrated decision support systems for interpreting CW forensic evidence.

Introduction

A typical forensic chemical agent detection scenario involves testing samples taken from the location of a suspected event, background samples from nearby locations that have not experienced the event, and "blank" samples prepared in the laboratory for contamination QA. Conclusions about the presence or absence of the agent are drawn from the observed pattern of "hits" on an analyte of interest, which may be the agent itself, or some proxy such as a precursor, impurity, or decomposition product. Variants of this procedure have been carried out in the Yellow Rain[1,2], Al Shifa[3], and Iraqi chemical weapon[4] (CW) investigations. In at least one case (Yellow Rain) claims were made that the results were the consequence of laboratory contamination.[5] In the Al Shifa affair, there was no report of background sampling results, which led to considerable public criticism.[3] In many such investigations circumstances often lead to the situation that the quantities of residual chemical signature are very low and there is a possibility that the analyte in question may also be present as a natural background.

There is a long-standing, if unpublicized, debate among both analysts and decision makers within the WMD forensics community about the evidentiary value of "hits" that are near the detection limit, or only appear sparsely among a set of questioned samples. One consequence of this uncertainty about how to determine the statistical significance of trace detection events is unnecessary conservatism – i.e. discounting trace detections unless they "clearly" exceed some threshold value that the analyst feels comfortable defending on intuitive grounds. The results of past investigations of alleged CW use suggest that the absence of more objective criteria may well compromise the value of such evidence in real cases in the future. Since forensic investigation of alleged CW events often take place under conditions where access is limited or risky, and where the collection of samples is necessarily opportunistic, one cannot generally address this issue by assuming that one may always revisit a site, or simply collect more samples.

To better understand the nature of this problem, consider an investigation where there are a certain number of "questioned samples" taken at a venue associated with a suspected activity involving a CW agent and "background samples" (control samples taken at places that are <u>not</u> associated with the activity in question.) Suppose that a few of the "questioned samples" have near-threshold hits, and the others are negative. Suppose further that all the "background samples" are negative. Are the few positive samples reportable evidence that the alleged activity involving the agent took place at the venue in question? In other words, how do we decide if our detection results are "*significant*"?

Most analysts and decision makers appreciate that it is <u>not</u> valid to conclude that the absence of hits among the "background samples" precludes the possibility that the "questioned sample" hits are due to background. If there is simply a low background presence of an analyte everywhere, it is not possible to categorically exclude the chance occurrence of a few positive "questioned sample" hits and no

"background sample hits" because the probability that any sample will turn up negative is very high. Clearly, to adequately interpret this kind of finding it would be useful to be able to estimate the probability of its occurrence under the hypotheses that the observed pattern of hits and negatives is simply due to presence of a low environmental background level of the analyte. Depending on the magnitude of this probability, a detection result may not be "significant" even if there are no background hits; conversely, a result may be "significant" even if there are some background hits.

In the absence of a transparent statistical significance-testing framework for trace detections, decisions about how positive or negative hits are reported depend largely on arbitrarily chosen criteria. When presented to decision makers or expert reviewers there is considerable danger that a given criterion may be questioned, injecting ambiguity and undermining confidence in any opinion rendered by the reporting laboratory. At worse, the criteria may be criticized as being either too conservative or not conservative enough, leading to controversy or even pressure to change a previous conclusion.

To address this I propose a simple null hypothesis significance testing (NHST) procedure for interpreting the pattern of "hits" among questioned samples, background samples, and contamination controls.[6] I am primarily interested in the case when all (or nearly all) of the detections are near threshold, i.e. trace detections. When concentrations of analytes are well above detection threshold, significance testing is far less necessary for obvious reasons. On the other hand, in cases where reliable concentration measurements are obtained for the tested samples, a natural NHST framework also exists, and I will discuss it briefly.

While the core technical argument of this paper is quite simple and requires only a few paragraphs to describe, the effective use of NHST in decision-making contexts involves some subtle conceptual issues. Most readers will probably recognize the basic statistical significance testing procedure from elementary statistics courses, where a P value is calculated and compared with some pre-set α value. Unfortunately this procedure is widely mis-understood and too often mis-used because its conceptual underpinnings are not generally taught correctly.[7] This leaves the naïve user of NHST procedures vulnerable to criticism, and may even lead to error in the way results are reported. Therefore I devote a substantial fraction of the paper to describing the conceptual underpinnings of the procedure and its roots in statistical decision theory as it applies to the forensic CW context.

Technical approach for trace detection NHST

For the background null hypothesis we will suppose that the analyte of interest is distributed continuously (but not necessarily uniformly) as a "background" in the environment, and our results can be explained entirely by this hypothesis. For the contamination null hypothesis we will assume that there is a certain probability of contamination of any sample under test, and all samples have the same vulnerability

to contamination. We will denote both of these hypotheses generically as H_0. When we wish to distinguish between background and contamination null hypotheses we will use B_0 and χ_0 respectively.

For this procedure to make sense we assume that there is some sort of standard extraction procedure (SOP) that is used on each environmental sample: a prescribed mass of sample is placed into a prescribed volume of extractant, and a prescribed sub-volume of the extractant is prepared for analysis using prescribed reagent volumes and procedures. Contamination control blanks are made up using a "standard environmental surrogate" (like a sterile soil sample) that is known not to contain the analyte, but can be extracted using the same procedures as the environmental (both questioned and background) samples.

The analysis procedure (also an SOP) takes a prescribed volume of the extract and performs an analysis for the concentration of the analyte in that extract, which we will denote C. Under the null hypothesis, for any sample drawn from the environment there is a probability $P(C|H_0)$ that the extract will contain the target analyte at concentration C.

Note that an estimate of $P(C|B_0)$ could be obtained *conceptually* by randomly sampling dirt, leaves, animals, or other samples from the relevant environment and performing the extraction/detection procedures on them. The histogram of observed C values would be a representation of $P(C|B_0)$. Similarly we could (conceptually) estimate $P(C|\chi_0)$ by running a large set of replicate blank samples mixed in with "live" samples. *We don't actually need to perform such sampling in order to perform the NHST procedure on a set of environmental samples and contamination controls*. For detections where the observed concentrations are close to (or below) the nominal detection threshold, we only need to know the number of questioned samples N_Q, background samples N_B, and contamination control samples N_C and the observed number of positive detections M_Q, M_B and M_χ to perform NHST. (When concentrations substantially larger than the detection threshold are measured, a standard non-parametric NHST method is described in the next section.) However, to *interpret* the P value generated by either of these tests we will need some additional information, as will be discussed further on.

Suppose that there is a threshold concentration for detection, C_{th}. Under the null hypothesis H_0, the probability that a sample will give a positive detection is:

$$P(+|H_0) = \int_{C_{th}}^{\infty} P(C|H_0)\, dC = \gamma \tag{1}$$

Similarly

$$P(-|H_0) = \int_{0}^{C_{th}} P(C|H_0)\, dC = 1 - \gamma \tag{2}$$

In equation (1) when $H_0 = B_0$, γ is estimated by the observed rate of positive detections among all the environmental samples (questioned + background) tested. Using M to represent the number of positive samples,

$$P(+|B_0) = \gamma_B \approx \frac{M_Q + M_B}{N_Q + N_B} \quad (3)$$

When $H_0 = \chi_0$, γ is estimated by the observed rate of positive detections among all samples (questioned + background + contamination control).

$$P(+|\chi_0) = \gamma_\chi \approx \frac{M_Q + M_B + M_\chi}{N_Q + N_B + N_\chi} \quad (4)$$

Given a constant rate of positive detections γ, the probability of observing M positive detections in N samples is given by

$$P(M, N) = \binom{N}{M} \gamma^M (1-\gamma)^{N-M} \quad (5)$$

Equation (5) can be used to calculate the probability of observing M detections in N samples under the null hypotheses for background and contamination by using equations (3) and (4) respectively to estimate the rate of "hits". We define two P values by noting that under the null hypothesis the partitioning of hits among the samples is assumed to be independent:

$$P_B = \binom{N_B}{M_B}\binom{N_Q}{M_Q} {\gamma_B}^{M_B + M_Q}(1 - \gamma_B)^{N_B - M_B + N_Q - M_Q} \quad (6)$$

$$P_\chi = \binom{N_B + N_Q}{M_B + M_Q}\binom{N_\chi}{M_\chi} {\gamma_\chi}^{M_B + M_Q + M_\chi}(1 - \gamma_\chi)^{N_B - M_B + N_Q - M_Q + N_\chi - M_\chi} \quad (7)$$

These are the probabilities of observing the specified number of positive detections among the relevant samples under each of the null hypotheses.
An example of P values calculated from equation (6) is shown in Table 1.

In the classical NHST procedure, the experimenter specifies a significance value α (e.g. in many contexts the value $\alpha = 0.05$ is used.) and if P is greater than α, the null hypothesis cannot be rejected at that significance level.[6] If the null hypothesis B_0 cannot be rejected ($P_B > \alpha$) then we cannot dismiss the notion that the detected signature is entirely due to a natural background with a probability of finding a concentration above detection threshold of γ. If the null hypothesis χ_0 cannot be rejected ($P_\chi > \alpha$) then we cannot dismiss the notion that the detected signature is entirely due to contamination with an average probability of contamination per sample of γ.

Table 1. P values for the case where there are no hits among N_B background samples and M_Q hits among ten questioned samples. Values in red are larger than 0.05.

	M_Q for N_Q = 10 samples									
N_B	1	2	3	4	5	6	7	8	9	10
1	0.350494	0.244419	0.190522	0.155177	0.128777	0.107314	0.088672	0.071446	0.054315	0.035049
2	0.319996	0.201882	0.140784	0.101159	0.072738	0.051270	0.034637	0.021677	0.011732	0.004486
3	0.294382	0.169562	0.106975	0.068767	0.043622	0.026641	0.015223	0.007790	0.003275	0.000891
4	0.272566	0.144429	0.083190	0.048380	0.027456	0.014793	0.007324	0.003170	0.001090	0.000230
5	0.253760	0.124499	0.065971	0.035028	0.017984	0.008668	0.003786	0.001420	0.000413	0.000071
6	0.237383	0.108429	0.053197	0.025985	0.012180	0.005311	0.002076	0.000687	0.000173	0.000025
7	0.222991	0.095282	0.043522	0.019683	0.008488	0.003379	0.001195	0.000354	0.000079	0.000010
8	0.210245	0.084389	0.036059	0.015183	0.006062	0.002220	0.000717	0.000192	0.000038	0.000004
9	0.198878	0.075264	0.030210	0.011899	0.004423	0.001500	0.000445	0.000109	0.000020	0.000002
10	0.188677	0.067543	0.025561	0.009458	0.003289	0.001038	0.000285	0.000064	0.000011	0.000001

Technical approach for NHST based on concentration measurements

When the test data consists of a set of concentrations, a framework for significance testing based on a standard non-parametric method can be used. Assume that we have obtained a set of concentrations $\{C_{q1}, C_{q2}, \cdots, C_{qN_q}\}$, $\{C_{b1}, C_{b2}, \cdots, C_{bN_b}\}$, and $\{C_{\chi 1}, C_{\chi 2}, \cdots, C_{\chi N_\chi}\}$, for the questioned, background, and contamination control samples respectively. If the null hypothesis B_0 is true, both the questioned samples and background samples have been drawn from the same distribution $P(C|B_0)$. Thus, a Kolmogorov-Smirnov (K-S) test based on the N_q questioned samples and N_b background samples can be used to calculate a P value.[8] Similarly, under the contamination null hypothesis χ_0, a K-S test based on comparing the $N_q + N_b$ environmental samples versus the set of N_χ contamination control samples can be used to calculate the P value.

When the test results are mixed, i.e some samples are below detection threshold entirely, some are above threshold but concentration estimates are uncertain, and some are well above threshold with precise concentrations, neither of the NHST approaches presented above is strictly applicable. However, a conservative approach would be to perform both kinds of test, *mutatis mutandis*, and choose the larger P value. A more statistically rigorous and satisfactory procedure remains to be developed.

Interpreting the P value

A great deal has been written about the use and abuse of P values in the NHST framework[9], and the procedure proposed above should not be utilized without understanding the role of significance testing as an element of a rational and transparent approach to reporting technical findings to decision makers. In this approach the selected significance level α is <u>not</u> arbitrarily chosen, as is sometimes suggested by statistics textbooks. Rather, it is a *derived* quantity, based on additional information about the alleged CW activity being investigated, prior

estimates of the probability that the event or activity in question occurred, and perceptions of the risk associated with reporting positive findings that might later be found to be incorrect. This section of the paper is intended to explain this often mis-understood aspect of NHST in the context of CW forensics.

To begin this discussion I note that detection results are almost always interpreted in the context of (at least) *two* competing hypotheses. The first is a null hypothesis as outlined above; the other is an alternative hypothesis – usually that the presence of the chemical agent is due to some CW related event. The "event" in question could be the alleged use of an agent or some incident involving the production, transport, or storage of an agent. The "significance" of NHST results can only be interpreted rationally when both a null and an alternative hypothesis are considered together.

The detection of the target analyte in a set of questioned, background and contamination control samples is evidence that can be used to choose among three distinct levels of reporting, epitomized by the following questions:[10]

(a) Which hypothesis (null or alternative) do my findings provide more support for?

(b) Which hypothesis do I now believe, in light of these findings?

(c) What decisions am I willing to make in light of this evidence?

All three questions come into play when setting the significance level α. Consider the background null hypothesis B_0. Let S_0 be the alternative hypothesis that the source of the analyte found in the questioned samples is a CW activity (use or production of a CW agent), and let $\boldsymbol{\Omega} = \{N_Q, M_Q, N_B, M_B\}$ represent the vector of observations on the questioned and background samples. According to Bayes theorem

$$\mathcal{O}(S_0|\Omega) = \frac{P(\Omega|S_0)}{P(\Omega|B_0)}\,\mathcal{O}(S_0) \tag{8}$$

where $\mathcal{O}(S_0)$ represents the prior odds that S_0 is true, and $\mathcal{O}(S_0|\Omega)$ represents the odds of S_0 being true in light of the observations Ω.[11] By definition, the denominator of this equation is the probability of observing the pattern of hits under the background null hypothesis, i.e. $P(\Omega|B_0) = P_B$, which is given by equation (6). The factor $P(\Omega|S_0)$ appearing in the numerator is the probability of observing the pattern of hits if S_0 were true. If $P_B < P(\Omega|S_0)$ then the observations Ω have increased the odds that the alternative hypothesis is true – in other words, our findings support the alternative hypothesis over the null hypothesis. $1/P_B$ is an upper-bound estimate of the likelihood ratio associated with the evidence provided by the observations. This answers question (a) above, but is the result "significant"?

To decide this we need to look at the meaning of the posterior odds $\mathcal{O}(S_0|\Omega)$ from the point of view of a fictitious case manager who is charged with reporting the results of the chemical analysis to decision makers above him. Normally, a decision to do something that carries risk requires that the odds of being correct are above some threshold. In this case we could imagine a case manager saying "I won't report this finding to my superiors unless the odds of being right are at least 10 to 1. If it's less, I'll say the test was not absolutely conclusive." The case manager's perception of the posterior odds is based partly on his estimate of the prior odds – i.e. how likely he judged S_0 to be *before* the chemical analysis results were in hand. For example, one could imagine him saying "Based on what we know about Iraqi chemical weapons doctrine I'd say that the odds are 1000 to 1 against them using CW in this situation." In this case, in order to convert $\mathcal{O}(S_0) = \frac{1}{1000}$ to $\mathcal{O}(S_0|\Omega) = \frac{10}{1}$ the value of P_B must be smaller than $1/10000$ because $P(\Omega|S_0)$ cannot be larger than 1. Thus, from the case manager's point of view, the analytical results are not "significant" with respect to question (c) unless the NHST P value is less than 10^{-4}.

An important observation is that higher P_B values might still lead the manager to believe S_0 to be more likely than not, answering question (b), yet not meet an appropriate standard of evidence to trigger the decision to report that the findings are conclusive proof of S_0 to higher ups. In reality this sort of calculation is seldom done so explicitly, but the reader should recognize the advantages of being able to think through the basis for choosing a certain α value, before making decisions about how results should be reported.

Another important issue raised by this analysis is that *the statistical significance of chemical detections also depends on details of the alternative hypothesis* S_0. Clearly, as the value of $P(\Omega|S_0)$ decreases, the upper limit for a "significant" P_B value becomes correspondingly smaller. Thus, complete interpretation of a "hit" also depends on whether that hit is consistent with our picture of the activity that we hypothesize occurred at the sampling site. Therefore it is important to report whether the hit is consistent with expectations derived from a plausible description of the activity as well as whether the null hypothesis can be rejected.

In practice, $P(\Omega|S_0)$ is estimated from models of the activity that estimate the amount of agent that would be deposited in plausible scenarios, from empirical estimates of signature decay rate, and from empirical data on signature collection efficiencies, extraction efficiencies, and detection limits. Ideally these are benchmarked by field experiments involving mock activities with surrogates. This sort of estimation of $P(\Omega|S_0)$ is also an essential element of good collection planning. Clearly a collection expedition, with its attendant risk and cost, would be unwarranted if the probability of observing signature above the detection threshold were extremely improbable.

Finally, it is important to recognize some limitations associated with the interpretation of P values derived from equations (6) and (7). These examples illustrate the fact that rejecting H_0 at a given significance level does not mean that we therefore "accept" S_0. Consider a case where ten questioned samples ($N_Q = 10$) and ten contamination control samples ($N_\chi = 10$) are analyzed along with only one background sample ($N_B = 1$), and suppose that only the background sample exhibits a "hit": i.e. $M_Q = M_\chi = 0$ and $M_B = 1$. The P value for the background NHST is 0.035, which leads to the (perhaps surprising) inference that we can reject B_0 at the 0.05 significance level. Thus, the test leads us to conclude that the uniform probability background hypothesis can be rejected in spite of the fact that *only the background sample tested positive*! As counterintuitive as this may seem, it correctly reflects the fact that the probability of observing no hits among ten questioned samples even though the apparent background hit rate is greater than 9% is very low. The P value for the contamination null hypothesis test, on the other hand, is nearly 0.2; thus the correct interpretation is indeed to reject B_0 – in favor of χ_0.

Along the same lines, note that equation (6) is symmetric with respect to exchange of (N_Q, M_Q) with (N_B, M_B). As a consequence, the P value for the case that we observe ten hits among ten questioned samples and zero hits among ten background samples ($P_B \approx 1 \times 10^{-6}$, clearly much smaller than 0.05) is equal to the P value for ten hits among ten background samples and zero hits among ten questioned samples. This correctly reflects the fact that when $\gamma \approx 0.5$ (ten hits among twenty samples) it would also be extremely unlikely to see ten background sample hits and no questioned sample hits if B_0 were true – it is much more likely that at least some of the questioned samples would exhibit hits as well. If this latter situation were to occur, the natural interpretation would be "none of the above" (P_χ is 0.00094) and some other hypothesis would be called for to explain the result.

Practical issues with the use of NHST in CW forensics

For the null hypothesis test to be considered a defensible basis for decisions, a number of additional conditions must be met.

- A level of significance α that triggers each reporting level should be agreed upon <u>prior</u> to testing, and be based on an analysis of prior probability and the required posterior probability as outlined above. *This prevents α from "sliding" after the analysis results become available.*

- The types and sizes of background collection samples must be similar to those of the test samples. The degree to which the types of samples must match is determined by the details of the background and alternative hypotheses being tested. For example, consider the samples analyzed by Black, et. al.[4] Suppose that one were concerned that the heat and pressure from ordinary bombs caused chemical reactions with naturally occurring soil

phosphate compounds that give rise to detectable traces of MPA and iPMPA. Then the background samples appropriately matched to soil samples from the craters of suspect nerve agent-dispersing bombs would be soil from craters known to have been caused by conventional bombs. Similarly, one of the criticisms of the mycotoxin analysis performed by Mirocha et. al. was that the background control samples they used were (apparently) not ones containing yellow spots - presumably bee feces - from locations where no alleged "yellow rain" incidents occurred.[5] Clearly there will be some cases where the most stringent efforts to match background and questioned samples may not satisfy a determined and motivated critic. Nonetheless, hypotheses about the potential sources for the background presence of the analyte should be actively sought out and taken into consideration when planning the collection.

- The background and questioned samples should be "anonymized" before they are given to the testing lab (e.g. labeled only with random number designators, and not associated with particular sampling sites).

- Blanks for contamination control should be prepared from surrogate materials like soil or wipes that enter the sample preparation laboratory in the same way that field samples do, and are handled and extracted the same way that the field samples are.

- The extraction method and detection assay should be well-calibrated, preferably for the actual sample materials that are collected. Moreover, the criteria for declaring a positive detection such as the magnitude of the detection threshold C_{th} should be transparently and objectively defined.

Generally it is prudent to draw one or more background samples in the vicinity of where the questioned samples are collected, at locations where CW signature analytes would not be expected under the alternative hypothesis S_0, but background would be expected, if it were present. Even where there is general agreement that no natural background exists, and the background hypothesis could be rejected without testing, negative background samples help to validate the testing procedure in other ways – e.g. as additional contamination controls.

What if it is not possible to collect an appropriate background sample in the same vicinity as the sample collection? (For example, a sample may come from an opportunistic collection by a native or NGO agent who may not think to collect a background sample.) In this case any decision about the significance of a "hit" in a collected sample in the absence of actual background samples will be based on a mental model of $P(\Omega|B_0)$ in the mind of one or more experts. For example, one could imagine an expert claiming "Based on my experience and intuition it is

unlikely that we would observe a signal as large as we did if it were only due to background." While expert *ipse dixit* is vulnerable to criticism based on the *availability* cognitive error, there are a variety of things that can stand in as mental proxies for actual background sampling data, such as historical background collections in other geographical regions, or the absence of any previous detection of the analyte in natural materials.

In general, however, the absence of a background sample changes a decision based on *risk* to one based on *ambiguity*, where only a range of possible probability values can be asserted. Ambiguity in reported technical findings can be expected to have well-documented aversion effects on decision making.[12] A formal analysis of this was performed by Tweney[13] whose results show that *at best* one could say that $\mathcal{O}(S_0|\Omega)$ lies between the original prior value $\mathcal{O}(S_0)$ and 1, but only if $P(\Omega|S_0) = 1$.

Interpreting and reporting the results of NHST calculations based on equations (6), (7) , or the K-S method must acknowledge that they do not address the likelihood of more complex hypotheses about background or contamination. For example, if calculations make the simple uniform probability background hypothesis very unlikely, a motivated critic could always invoke the more complex hypothesis that there is some source of natural background that happens to be localized near the site where the questioned samples were collected but is (miraculously?) absent where the background samples were collected. Now, however, perceptions about where the burden of proof lies often shift to those putting forward the more complex hypothesis. The use of NHST as proposed in this paper can be viewed as a variant of Occam's razor: only if the simplest hypothesis can be rejected is it sensible to consider more complicated hypotheses, and S_0 is often the simplest remaining hypothesis.

NHST in an integrated decision structure

Ideally NHST is used as a component of a more general integrated structure for interpreting and reporting forensic findings. To make the most accurate estimates of confidence, this structure must capture uncertainties about the details of the CW activity that is alleged to have occurred, the conditions under which the signature was produced and propagated to the sampling location, conditions that might cause the signature to decay with time, and uncertain factors in collection and sample handling. In addition it must be able to integrate other types of information such as eyewitness reporting and other types of intelligence. One natural structure is the Bayesian Network, which has gained popularity in similar applications over the last decade.[14] A simple BN that has the NHST calculations embedded in it is shown in Fig. 2.

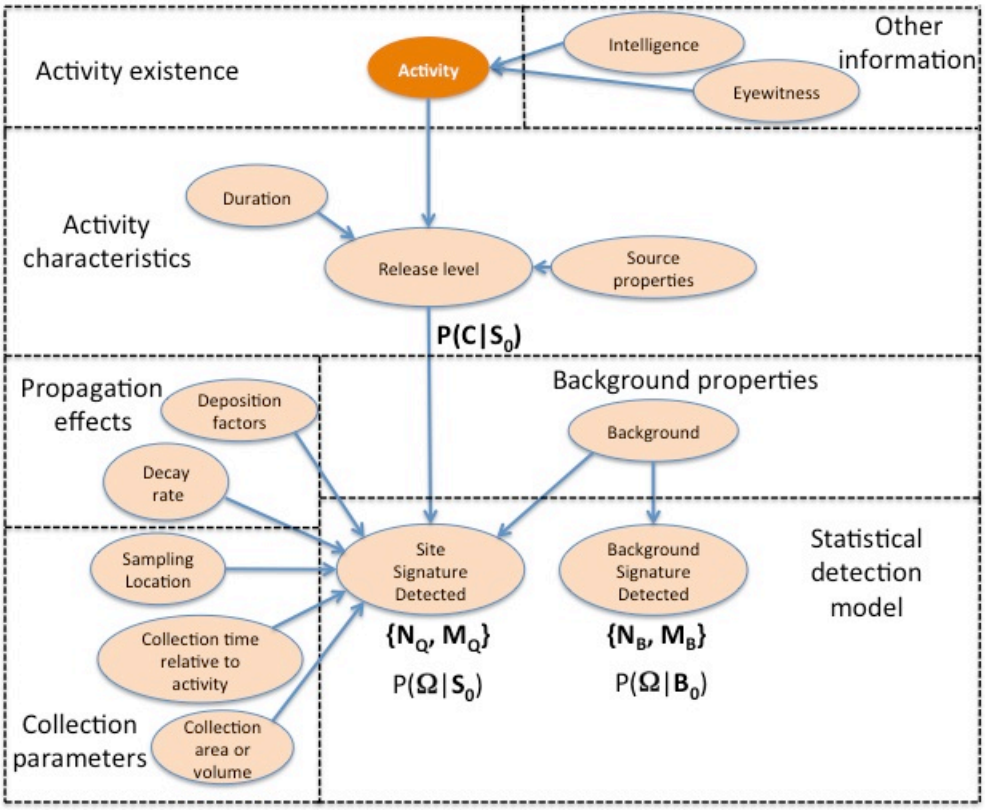

Figure 2. A notional Bayesian network to assess the likelihood that a hypothesized activity exists, or existed. Observed detections would be used to instantiate the detection nodes.

Another approach views NHST in the framework of the "analysis of alternative hypotheses" methodology.[15] This methodology has been reviewed in many contexts, and basically consists of 5 steps:

(1) Identify at least one alternative hypothesis in addition to your primary one; make sure your individual pieces of evidence are truly independent of each other.

(2) Consider your evidence in light of each hypothesis and assess how likely it would be to obtain those results if that hypothesis were true.

(3) Prepare a matrix with items of evidence down and hypotheses across; place your likelihood estimates in the corresponding boxes; append a last row that contains estimates of how likely you would estimate each hypothesis in the *absence* of all the evidence you have listed.

(4) Rank the evidence items according to their ability to differentiate among hypotheses; eliminate those items of evidence that have nearly the same likelihoods under all hypotheses.

(5) Rank the hypotheses according to the least number of low likelihood entries in

their corresponding column (there may be ties).

For significance testing we restrict our attention to three hypotheses, S_0 and B_0 and χ_0, leading to an ACH matrix like the one shown in Fig. 3.

Analysis of competing hypotheses		Hypotheses		
		H_1	H_2	H_3
Evidence	E_1	$P(E_1 \mid H_1)$	$P(E_1 \mid H_2)$	$P(E_1 \mid H_3)$
	E_2	$P(E_2 \mid H_1)$	$P(E_2 \mid H_2)$	$P(E_2 \mid H_3)$
	P_0	$P(H_1)$	$P(H_2)$	$P(H_3)$

Significance test for background		Hypotheses		
		S_0	B_0	χ_0
Evidence	Ω	$P(\Omega \mid S_0)$	$P(\Omega \mid B_0)$	$P(\Omega \mid \chi_0)$
	EWT	$P(EWT \mid S_0)$	$P(EWT \mid B_0)$	$P(EWT \mid \chi_0)$
	MED	$P(MED \mid S_0)$	$P(MED \mid B_0)$	$P(MED \mid \chi_0)$
	P_{prior}	$P(S_0)$	$P(B_0)$	$P(\chi_0)$

Figure 3. A representation of the Analysis of Competing Hypotheses matrix.

Here, EWT represents evidence from eye-witness testimony about how and where the agent was dispersed and the magnitude of the attack; MED represents evidence derived from medical studies of alleged victims and non-victim control patients. Clearly the likelihood of observing particular EWT and MED evidence would depend strongly on the details of the hypothesis S_0, and MED might also depend on the details of B_0 – since naturally occurring background chemicals might have observable effects on health. It seems reasonable to assume that the likelihood of observing both eye-witness details and medical evidence about a CW incident would be expected to be low if laboratory contamination were the sole explanation of positive test results.

Summary and conclusions

The null hypothesis significance tests outlined in this paper are a simple but useful tool for guiding the interpretation of detection "hits" in field samples. It is important to note that a result may not be significant even if there are no background hits; conversely, a result may be significant even if there are some background hits. Similarly with regard to contamination, detection results may not be significant even if there are no contamination hits and a result may be significant even if there are some hits among the negative control samples.

Significance levels (α) are not set arbitrarily – they should be based on a consideration of prior odds and the tolerable level of risk that a wrong call may incur. In addition to determining the P value under the null hypothesis (background or contamination), it is important to ascertain if the observed signature levels are consistent with the source (alternate) hypothesis.

Finally, NHST procedures transparently produce likelihood estimates that can be used in more general decision support structures such as Bayesian Nets and the Analysis of Competing Hypotheses. They are useful tools both for planning collection activities, estimating the potential value of such activities prior to execution, for evaluating the results, and for expressing the findings.

The framework for significance testing brings into sharp relief the issue of ambiguity in the reporting of test results in forensics. When probabilities can be estimated it facilitates decisions because risk is well defined, whereas if one cannot estimate probabilities then decisions – even decisions about the proper way to report the findings – become difficult. All source analysts and other consumers of forensic findings should demand that findings come with rigorous, transparent, and traceable assessments of uncertainty. In this sense the primary function of background and contamination controls is to help *define* risk by allowing us to estimate probabilities – controls *per se* do not eliminate risk.

References

1. Rosen RT and Rosen JD, "Presence of Four Fusarium Mycotoxins and Synthetic Material in Yellow Rain", Biomedical Mass Spectrometry, 1982; **9**(10): 443-450.

2. Mirocha CJ, Pawlosky RA, Chatterjee K, Watson S, and Hayes W, "Analysis for Fusarium Toxins in Various Samples Implicated in Biological Warfare in Southeast Asia", J. Assoc. Official Anal. Chemists 1983; **66**(6): 1485-1499.

3. Barletta M, "Chemical Weapons in the Sudan: Allegations and Evidence", The Nonproliferation Review, Fall 1998; pp 115-136.

4. Black RM, Clarke RJ, Read RW, and Reid MTJ, "Application of gas chromatography – mass spectrometry and gas chromatography – tandem mass spectrometry to the analysis of chemical warfare samples, found to contain residues of the nerve agent sarin, sulphur mustard and their degradation products", Journal of Chromatography A 1994; **662**: 301-321.

5. Meselson MS and Robinson JP, "The Yellow Rain Affair: Lessons from a Discredited Allegation", in Terrorism, War, or Disease, Clunan AL, Lavoy PR, and Martin SB, eds. (Stanford University Press, Stanford California, 2008).

6. Descriptions of this procedure can be found in many general statistics textbooks; See for example Motulsky H, *Intuitive Biostatistics*, (Oxford University Press, Ne York, 1995).

7. Lecoutre MP Poitevineau J, Lecoutre B, ""Even statisticians are not immune to misinterpretation of Null Hypothesis Significance Tests", International Journal of Psychology 2003; **38**: 37-45.

8. Corder GW and Foreman DL, Nonparametric Statistics for Non-Statisticians, a Step by Step Approach, (J Wiley & Sons, New Jersey, 2009).

9. D'Errico G, "Issues in Significance Testing", Measurement 2009; **42**: 1478-1481.

10. This discussion draws on Richard Royall's monograph Statistical Evidence: A Likelihood Paradigm (Chapman and Hall/CRC Monographs on Statistics and Applied Probability, 1997). This book also contains an excellent discussion of problems with NHST. My treatment can be recognized by experts as a sort of "Bayes Lite", and is designed to appeal to scientists whose only prior statistics exposure is to classical NHST procedures.

11. In writing Bayes's equation this way I am implicitly assuming that there are only two hypotheses that can explain the evidence. The prior odds of S_0 are then inversely related to the prior odds for B_0, because S_0 is taken to be equivalent to

"not-B_0." This is done for pedagogical convenience, but is not rigorous because not-B_0 could include hypotheses other than S_0, for example more complex background hypotheses. However, if S_0 is considered the most likely alternative to B_0 the treatment is approximately correct. Note also that only the K-S framework provided in this paper allows the alternative hypothesis S_0 to include an added natural background term, where B_0 is then the "background only" hypothesis.

12. Ghosh D and Ray MR, "Risk, Ambiguity, and Decision Choice: Some Additional Evidence", Decision Sciences 1997; **28**(1): 81-91; In older literature ambiguity is sometimes called "uncertainty"; modern terminology uses uncertainty to denote cases where probabilities can be defined and ambiguity the situation where they cannot.

13. Tweney RD, Doherty ME, and Kleiter GD, "The pseudodiagnosticity trap: Should participants consider alternative hypotheses?", Thinking and Reasoning 2010; **16**(4): 332-345.

14. Taroni F, Aitken C, Garbolino P, and Biederman A, Bayesian Networks and Probabilistic Inference in Forensic Science, (J. Wiley & Sons, 2006).

15. U.S. Government, "A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis", March 2009.